La résolution d'un problème de multicolinéarité au sein des études portant sur les déterminants d'une publication volontaire d'informations : proposition d'un algorithme de décision simplifié basé sur les indicateurs de Belsley, Kuh et Welsch (1980)

Marc de Bourmont

La résolution d'un problème de multicolinéarité au sein des études portant sur les déterminants d'une publication volontaire d'informations : proposition d'un algorithme de décision simplifié basé sur les indicateurs de Belsley, Kuh et Welsch (1980)

Marc de Bourmont
Enseignant-Chercheur, Rouen Business School
Courriel : mdb@rouenbs.fr

Résumé :

De nombreuses études ont été réalisées sur le thème des déterminants d'une publication volontaire d'informations et la plupart des auteurs de ces études recourent à l'estimation d'un modèle de régression linéaire classique (RLC). Cette technique aboutit en effet, lorsque les contraintes nécessaires à son application sont respectées, à l'utilisation des meilleurs estimateurs linéaires sans biais. Une absence de multicolinéarité constituant l'une de ces contraintes, l'objet de cet article est 1/de montrer l'insuffisance relative des outils de diagnostic généralement utilisés 2/ de démontrer les avantages et la supériorité des indicateurs de Belsley, Kuh et Welsch (1980). A partir de ces indicateurs, cet article propose un algorithme de décision simplifié qui est appliqué d'une façon illustrative au sein d'une étude empirique portant sur les déterminants d'une publication volontaire d'informations sur les activités de recherche et développement.

Mots clés :
Régression linéaire, Multicolinéarité, Variance Inflation Factor, VIF, Matrice des Corrélations.

Abstract :

Many disclosure and determinants studies have been conducted, in which the authors most often estimate a classical linear regression model (CLR). This technique is the best technique in order to estimate linear relationships between a dependent variable and some independent variables, when the constraints that apply to such a model are respected. The lack of a multicolinearity problem being one of these constraints, the aim of this article is 1/ to show that the usual tools that are used in order to detect and remedy this problem are sometimes not sufficient 2/ to demonstrate the superiority of the indicators that were proposed by Besley, Kuh and Welsch (1980). From these indicators, an algorithm is proposed in this article which is then applied in an illustrative way in a determinant and disclosure study. We then show that this algorithm allows circumventing a multicolinearity problem.

***Key words** :*
Linear Regression, Multicolinearity, Variance Inflation Factor, VIF, Correlation Matrix

Introduction

De nombreuses études ont été réalisées sur le thème des déterminants d'une publication volontaire d'informations au sein des rapports annuels d'entreprises (pour une revue de la littérature, le lecteur pourra se référer aux articles d'Ahmed et Courtis (1999), Pourtier (2004), Chavent et al. (2006) et Garcia-Meca et Sanchez-Balesta (2010)).

D'une façon générale, la publication volontaire d'informations peut être définie comme « la publication d'informations non requises (par les normes comptables ou les lois en vigueur), représentant en cela un choix laissé à la discrétion des dirigeants d'entreprises afin de leur permettre de fournir des informations de nature comptable ou de toute autre nature qui pourraient s'avérer utiles à la prise de décision des utilisateurs de ces informations » (Meek et al., 1995a, p.255)[1]. Et, dans le contexte de ces études, le mot « déterminant » peut s'entendre comme un synonyme de « facteur explicatif ».

L'objet de ce champ de recherche est de tenter d'expliquer des variations concernant les niveaux de publication volontaire recensés à partir d'un dépouillement des rapports annuels d'entreprises par une ou plusieurs caractéristiques (déterminants) de ces entreprises (par exemple : la taille, le secteur d'activité, la rentabilité des entreprises observées…).

L'étude théorique du comportement de publication (ou au contraire de rétention) volontaire d'informations conduit à l'examen de plusieurs cadres d'analyse interdépendants, selon lesquels la décision de publier ou non une information de nature volontaire au sein du rapport annuel, va résulter de la comparaison qui sera faite entre :

- d'une part, les « économies de coûts » qui pourraient résulter de cette publication. Selon les préceptes de la théorie politico-contractuelle (Watts, 1977 ; Watts et Zimmerman, 1978, 1986), les dirigeants publieraient ainsi des informations sur un mode volontaire à propos de leur société afin de diminuer, d'une part, les coûts (dits d'agence) liés aux conflits d'intérêt pouvant exister entre les diverses parties contractantes de l'entreprise et, d'autre part, les coûts (dits politiques) qui pourraient être liés à une taxation par les pouvoirs publics en cas de visibilité trop importante de leur entreprise (à la suite de profits élevés notamment). La théorie du signal (Akerlof, 1970) avance une explication supplémentaire. Les dirigeants choisiraient de publier des informations de nature volontaire pour signaler aux marchés financiers la qualité de l'entreprise qu'ils dirigent, afin d'obtenir un coût du capital plus avantageux pour leur société (cas des entreprises en bonne santé) ou pour éviter une diminution trop importante du cours de l'action de leur entreprise (phénomène d'anti-sélection en cas d'absence de visibilité sur les perspectives financières de l'entreprise).

- et, d'autre part, les coûts qui pourraient inversement résulter de cette publication. Une publication volontaire d'informations est en effet susceptible de générer d'une part des coûts

[1] Voluntary disclosures can be defined as « *disclosures in excess of requirements, representing free choices on the part of company managements to provide accounting and other information deemed relevant to the decision needs of users of this information.* »

« matériels » supplémentaires (dits coûts directs) et, d'autre part, des coûts dits « indirects » (Verrecchia, 1983) : les concurrents d'une entreprise - ou tout autre groupe de pression - pourraient en effet utiliser l'information publiée d'une façon qui serait néfaste à cette entreprise. L'existence de ces deux types de coûts pourrait donc inciter les dirigeants à ne pas rendre publique l'information privée dont ils disposent à propos de leur entreprise.

Les déterminants étudiés sont ainsi et le plus souvent représentatifs de l'importance de ces différents coûts et les études appartenant à ce champ de recherche sont en général menées en trois temps :

- les auteurs opèrent, premièrement, un dépouillement des rapports annuels qui conduit à disposer d'un niveau / score / indice de publication volontaire d'informations pour chacune des entreprises observées. Ces niveaux constituent alors la variable à expliquer.

- ils énoncent, ensuite, des hypothèses reliant la variable à expliquer à chacun des déterminants (variables explicatives) qu'ils ont choisi d'étudier.

- enfin, des tests statistiques sont réalisés, qui permettent de savoir si les hypothèses posées sont validées (et donc si les théories justificatives d'une publication / rétention volontaire d'informations sont confirmées).

La plupart des auteurs recourent lors de cette troisième étape à la réalisation d'un modèle de régression linéaire classique (RLC), où les paramètres sont estimés à partir de la méthode des moindres carrés ordinaires (MCO). Un modèle RLC peut être présenté ainsi :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \mu_i$$

où :

- Y_i est la variable dépendante (ou variable à expliquer)

- β_0 est la constante de régression

- X_{1i}, X_{2i}, X_{3i}, …, X_{kj} sont les variables indépendantes (ou variables explicatives, ou régresseurs)

- β_1, β_2, β_3, …,β_k sont les coefficients de régression

- et μ_i est le terme d'erreur (ou résidu(s), ou erreur(s))

La popularité de cette technique s'explique par le fait qu'elle est simple d'utilisation et qu'elle conduit à un recours aux meilleurs estimateurs linéaires sans biais (« Best Linear Unbiased Estimators », dits « BLUE »), les estimateurs des MCO, dès lors qu'un certain nombre d'hypothèses (ou conditions) sont respectées. Ces conditions sont les suivantes (Gujarati, 2004, p.340):

> « Hypothèse 1. Le nombre d'observations doit être supérieur à celui des régresseurs.
>
> Hypothèse 2. Le modèle est linéaire quant aux paramètres.
>
> Hypothèse 3. Le modèle de régression est correctement spécifié.
>
> Hypothèse 4. Les valeurs des régresseurs, les *X*, sont fixes dans des échantillons répétés.

Hypothèse 5. Si les *X* sont stochastiques, le terme d'erreur et les *X* sont indépendants ou, au minimum, non corrélés.

Hypothèse 6. Pour des *X* donnés, la valeur moyenne de l'erreur μ_i est nulle.

Hypothèse 7. Pour des *X* donnés, la variance de μ_i est constante ou homoscédastique.

Hypothèse 8. Le terme d'erreur stochastique μ_i présente une distribution normale.

Hypothèse 9. Pour des *X* donnés, il n'y a pas d'autocorrélation des erreurs.

Hypothèse 10. Il doit exister une stabilité suffisante dans les valeurs prises par les régresseurs.

Hypothèse 11. Il n'existe pas de relation linéaire exacte (multicolinéarité) entre les régresseurs. » [2]

Lorsque ces hypothèses sont validées, on montre en effet que les estimateurs des MCO sont (Bourbonnais, 2005) :

- « linéaires, c'est-à-dire une fonction linéaire d'une variable aléatoire telle que la variable dépendante *Y* du modèle de régression.

- sans biais. Leur moyenne, ou valeur espérée E (β), est égale à la « vraie » valeur de β.

- et performants (efficients). Ils présentent une variance minimale dans toutes les possibilités des estimateurs linéaires sans biais. »

On peut alors tester des hypothèses sur la valeur vraie des coefficients β, connaitre leur niveau de signification statistique et interpréter de façon non biaisée les différentes statistiques issues de l'estimation de la régression. Ces statistiques, le plus souvent automatiquement fournies par les logiciels économétriques actuels, sont les suivantes :

- le R^2, qui est une estimation de la proportion de la variable dépendante qui est expliquée par la régression ;

- l'erreur standard des résidus (la valeur de l'erreur standard de l'estimation) ;

- l'ANOVA (contraction de « ANalysis Of VAriance ») de la régression, qui a pour but de tester s'il existe une relation linéaire entre les variables, en formant un rapport *F* du carré moyen de la régression sur le carré moyen des résidus. L'ANOVA fournit donc la valeur et le niveau de signification statistique de *F*, ces deux éléments permettant de statuer quant au niveau de significativité globale du modèle ;

- les valeurs des coefficients β ainsi que les valeurs des écarts types estimés par les estimateurs des coefficients (Erreurs standard ou « Standard Errors » en anglais) et les valeurs de l'intervalle de confiance pour chaque coefficient ;

[2] Nous tenons ici à informer le lecteur que les propos contenus au sein de cet article sont relatifs aux modèles de régression en coupe instantanée (transversale) au sein desquels la variable à expliquer est de nature quantitative. Les indicateurs de BKW peuvent cependant être appliqués à d'autres types de modèles (et notamment les modèles de régression logistique, régressions poolées et/ou régressions en données de panel) mais les conditions de validité de ces modèles peuvent alors différer (pour partie) des onze conditions évoquées ici.

- les valeurs des statistiques *t*, qui permettent de savoir, pour chacune des variables explicatives retenues dans le modèle de régression, si elle contribue à la puissance prédictive de l'équation de régression, et leur niveau de signification statistique *p*.

Avant d'estimer un modèle RLC, tout utilisateur averti doit donc tester du respect de ces conditions. Les techniques pour ce faire sont en général exposées au sein de tout ouvrage d'Econométrie, le non-respect de l'une des onze hypothèses évoquées conduisant à l'impossibilité de recourir à un modèle RLC et, le plus souvent, à la nécessité du choix d'un autre type d'estimateurs que les MCO. Les dix premières hypothèses peuvent être facilement testées et les mesures à prendre en cas de non-respect de celles-ci sont connues[3].

Le non respect de l'hypothèse 11 est en revanche plus problématique et il s'avère qu'il se pose fréquemment au sein des recherches portant sur les déterminants d'une publication volontaire d'informations, où le nombre de déterminants étudiés peut être important. Des solutions ont été proposées, dont nous verrons qu'elles ne sont que relativement satisfaisantes.

L'objet de cet article est donc de présenter les techniques principales qui existent à ce jour pour 1/ détecter un phénomène de multicolinéarité 2/ résoudre les problèmes liés à l'existence d'un tel phénomène. A partir d'une étude empirique portant sur les déterminants d'une publication volontaire d'informations sur les activités de recherche et développement, il sera montré en quoi les indicateurs proposés par Belsley, Kuh et Welsch (1980) - ci-après BKW - constituent des outils de diagnostic efficients de la multicolinéarité et un algorithme de décision simplifié basé sur ces indicateurs sera présenté afin de permettre de maitriser les problèmes liés à l'existence d'un tel phénomène.

Au sein d'une première section, les problèmes engendrés par la présence d'un phénomène de multicolinéarité sont énoncés et les solutions habituelles permettant de résoudre ces problèmes sont décrites. La section 2 permet quant à elle de présenter les indicateurs proposés par BKW (1980) et l'algorithme de décision simplifié. Une troisième section, enfin, énonce les modalités et les résultats de l'étude empirique illustrative.

1. Les problèmes engendrés par l'existence d'un phénomène de multicolinéarité

Lorsqu'il existe une relation linéaire exacte entre deux ou plusieurs régresseurs, une équation de régression linéaire devient insoluble. Dans ce cas, il convient de ne conserver pour l'analyse que l'une des variables explicatives parmi celles qui sont parfaitement corrélées entre elles. Ce

[3] Cet article étant prioritairement consacré à l'étude des problèmes liés à l'existence d'un phénomène de multicolinéarité, les différents tests statistiques permettant de valider les dix premières hypothèses et les solutions en cas de non-respect de celles-ci ne seront pas ici décrits en détail. Néanmoins, les principaux tests utilisés sous STATA 11, qui est le logiciel auquel il a été recouru dans le cadre de cet article, sont indiqués en Annexe 1, étant entendu que cette Annexe présente les tests les plus couramment utilisés et ne détaille pas les lignes de commande à entrer sous STATA pour ce qui est des solutions. Nous invitons donc le lecteur à consulter plus avant des ouvrages d'Econométrie si telle ou telle de ces dix hypothèses n'était pas respectée.

traitement ne pose pas de problème particulier, dans la mesure où les variables concernées seront le plus souvent susceptibles de représenter une seule et même « réalité ».

L'existence d'une relation linéaire importante, quoiqu'imparfaite, entre deux ou plusieurs régresseurs est susceptible en revanche d'engendrer des problèmes statistiques plus subtils. En effet, bien qu'il existe une corrélation entre ces variables, il n'est, au contraire du cas précédent, pas évident que ces variables représentent une même « réalité ». L'expérimentateur souhaitera donc conserver toutes les variables explicatives dans le cadre de son analyse pour que celle-ci soit la plus riche possible. Mais d'un autre côté, l'inclusion de l'ensemble des variables est susceptible de générer d'importants problèmes liés à la colinéarité imparfaite existant entre elles, problèmes dont les symptômes les plus courants sont les suivants :

- des erreurs standard dont les niveaux seront importants pour les variables concernées, avec des statistiques *t* très faibles au contraire pour ces variables ;

- des changements de signe inattendus concernant le sens de la relation existant entre la variable dépendante étudiée et les variables explicatives concernées, ou une amplitude exacerbée des coefficients de régression pour les variables indépendantes concernées ;

- des coefficients de régression non significatifs alors que l'analyse conduit à l'obtention d'un R^2 élevé.

Plusieurs « outils » statistiques ont été créés en vue :

- premièrement, de déceler un phénomène de multicolinéarité ;

- deuxièmement, de remédier à ce phénomène dans le cas où celui-ci devra être traité.

Les outils permettant de détecter un phénomène de multicolinéarité

Deux techniques sont habituellement utilisées : la réalisation d'une matrice des corrélations et le calcul des VIFs (« Variance Inflation Factors »).

La réalisation d'une matrice des corrélations permet une analyse deux à deux des corrélations entre variables explicatives. Il est d'usage de considérer que l'obtention de coefficients de corrélations supérieurs à 0,5 est révélatrice d'un problème de multicolinéarité entre les variables concernées.

La seconde solution consiste à régresser chacune des variables explicatives sur les autres. En effectuant le calcul $(1 - R^2)$ à partir de chacune des régressions opérées, il est alors possible de savoir quelle part de la variance d'une variable explicative est indépendante des autres variables explicatives, le calcul $(1/(1-R^2))$ permettant alors d'obtenir une statistique « VIF » pour chaque variable. Sous STATA, les VIFs sont obtenus en utilisant la commande post-régression « **vif** ». Un problème de multicolinéarité est relevé dès lors qu'un VIF présente une valeur supérieure ou égale à 10 et/ou lorsque la moyenne des VIFs est supérieure ou égale à 2 (Chatterjee, Hadi et Price, 2000). Si aucune de ces deux valeurs n'est atteinte, l'impact de la multicolinéarité n'est, selon ces auteurs, pas inquiétant et toutes les variables explicatives peuvent donc être conservées

pour l'analyse, cette dernière n'étant alors pas « faussée » de manière rédhibitoire par le niveau de multicolinéarité existant. Si, au contraire, ces valeurs étaient atteintes, le problème de multicolinéarité devrait alors être traité par l'expérimentateur.

Les outils permettant de traiter un problème de multicolinéarité

Plusieurs outils sont habituellement utilisés : la réalisation de modèles alternatifs simplifiés, d'un modèle de régression pas-à-pas ou d'une analyse factorielle.

La réalisation de modèles alternatifs simplifiés est une méthode simple dans son exposé. Elle consiste à estimer toutes les combinaisons possibles (soient 2^p-1 possibilités, p étant le nombre de régresseurs). Cette méthode permet alors d'identifier les relations existant entre la variable à expliquer et les variables explicatives en contournant le problème de multicolinéarité.

Dans le cas de l'estimation d'une régression pas à pas (« stepwise regression »), il s'agit de retenir le modèle qui sera composé des variables explicatives :

- les plus corrélées avec la variable à expliquer ;

- les moins corrélées entre elles.

Cette procédure consiste à introduire les régresseurs un par un dans l'équation de régression et à ne conserver que ceux qui sont les plus significativement associés avec la variable à expliquer. Les autres variables explicatives sont alors « éliminées » de la régression.

Pour ce qui concerne la réalisation d'une analyse factorielle, et dans la mesure où deux ou plusieurs régresseurs sont fortement « corrélés » entre eux, une solution peut consister à « factoriser » ces variables, c'est-à-dire à regrouper les variables corrélées entre elles de façon à ce qu'elles ne forment qu'une seule et même variable qui sera dénommée dans ce cas un « facteur ». La factorisation a un double intérêt : d'une part, la disparition du problème de multicolinéarité (les facteurs extraits étant orthogonaux) et, d'autre part, la conservation de l'ensemble des variables dans l'analyse.

2. L'insuffisance des outils usuels : présentation des indicateurs de BKW (1980) et de l'algorithme de décision simplifié

Malgré leur intérêt, les outils qui viennent d'être présentés souffrent de nombreuses limites, dont certaines sont évoquées au sein de l'article de BKW (1980), et qui sont résumées ci-après.

S'agissant des outils de détection :

- la réalisation d'une matrice des corrélations n'inclut pas l'étude d'une possible colinéarité entre les variables explicatives et la constante de régression. Or l'existence d'une telle relation peut fausser les résultats obtenus. En outre, le critère lié à l'obtention d'un ou plusieurs coefficients de corrélation de valeur supérieure à 0,5 s'est parfois révélé insuffisant, un phénomène de

multicolinéarité ayant pu être constaté lorsque les coefficients de corrélation présentaient une valeur comprise entre 0,3 et 0,5.

- le calcul des VIFs est également problématique. Premièrement, et tel qu'il est opéré au moyen de la plupart des logiciels, ce calcul ne prend pas non plus en compte l'existence d'une éventuelle colinéarité entre les régresseurs et la constante. Sous STATA par exemple, les VIFs sont calculés après que les valeurs des variables explicatives aient été « centrées », ce qui ne permet pas de prendre en compte la constante dans l'analyse. L'ajout d'une option « uncentered » à la fonction « vif » (**« vif, uncentered »**) permet de combler cette lacune. Mais les VIFs présentent en effet une autre lacune : s'ils renseignent sur de possibles relations colinéaires, ils ne prennent pas en compte l'existence de « dépendances proches » (voir ci-dessous) pouvant avoir un impact important sur les paramètres estimés.

- enfin, et pour ce qui concerne ces deux outils, ils ne conduisent qu'à pouvoir diagnostiquer la présence d'un phénomène de multicolinéarité, mais ils ne permettent, ni l'un ni l'autre, d'évaluer la portée exacte de ce phénomène sur les résultats obtenus. Ils ne répondent donc que partiellement au problème.

Et, s'agissant des méthodes de résolution :

- la réalisation de modèles alternatifs simplifiés peut s'avérer fastidieuse lorsque le nombre de régresseurs est important.

- une régression pas-à-pas permet d'extraire les variables explicatives qui sont les plus significatives mais le problème de multicolinéarité n'est pas solutionné pour les autres variables.

- enfin, une factorisation ne permet pas d'étudier l'impact de chacun des régresseurs sur la variable dépendante, les variables explicatives se présentant alors sous la forme de facteurs devant être interprétés comme tels.

Pour remédier à ces lacunes, BKW (1980) proposent une autre méthode, connue sous le nom « d'indicateurs de BKW ».

La méthode consiste à calculer, dans un premier temps, deux types d'indicateurs, dénommés respectivement « indices de conditionnement » (« condition numbers ») et « proportions de décomposition des variances » (« variance-decomposition proportions »). La première statistique est indicatrice du niveau de multicolinéarité général existant entre chacun des régresseurs et les autres. La seconde permet de calculer non pas les coefficients de corrélation existant entre les variables explicatives considérées deux à deux mais ce que les auteurs dénomment les « proportions de décomposition des variances ». Cette statistique emploie tout à la fois la technique des VIFs et celle d'une analyse en composantes principales. Sous STATA, un tableau des indices de conditionnement et des proportions de décomposition des variances (dit « tableau de décomposition des variances ») est obtenu en exécutant la commande post-régression « **coldiag2** » [4]. Le tableau obtenu comprend autant de lignes et de colonnes qu'il existe de

[4] L'estimation des indices de conditionnement et des proportions de décomposition des variances doit être réalisée une fois que les variables explicatives ont été normées à 1 (voir Erkel-Rousse, 1995, p.23). La commande « **coldiag2** » sous STATA opère automatiquement cette transformation, ce qui n'était pas le cas de la commande antérieure « **colldiag** ». On préférera donc l'utilisation de la commande « **coldiag2** ». En outre, un débat existe sur le

variables explicatives (constante incluse). Il permet de présenter, au sein d'une première colonne, un indice de conditionnement pour chacune des variables explicatives et, sur la ligne correspondant à chaque indice de conditionnement, les proportions de décomposition des variances relatives à l'ensemble de ces variables » [5].

Une situation où un indice de conditionnement ressort avec une valeur supérieure à 30 et présente sur la ligne correspondante au moins deux proportions de décomposition des variances supérieures à la valeur 0,5 est appelée « dépendance proche » et conduit à constater l'existence d'un phénomène de multicolinéarité entre les régresseurs concernés.

Le dénombrement des dépendances proches étant essentiel pour la suite, il convient d'opérer une différence entre les dépendances proches « dominantes » (« dominating near dependencies » ») et les dépendances proches « concurrentes » (« competing near dependencies »).

La règle d'interprétation de BKW (1980) étant présentée, comme le note Erkel-Rousse (1995), « sous une forme très peu synthétique », il a été choisi d'adopter ici la règle d'interprétation « RI » édictée par cette auteure pour quantifier le nombre de dépendances proches (cf Erkel-Rousse, 1995, p.26).

« 1/ En pratique, le seuil à partir duquel un indice de conditionnement peut être considéré comme élevé se situe à 30 environ mais des valeurs un peu inférieures (à partir de la vingtaine) sont ambigües, et doivent donc également attirer l'attention[6].

2/ Supposons que *p* indices de conditionnement exactement soient élevés (>20). Alors :

- les indices de conditionnement présentant une valeur inférieure à 20 ne sont pas pris en compte.

- les indices de conditionnement présentant une valeur supérieure à 20 et présentant, pour chaque vecteur *X* et sur les lignes correspondant à ces indices et sur les lignes directement situées au-dessus dès lors que la valeur des indices de conditionnement correspondants dépasse 20, une somme des proportions de décomposition des variances inférieure à 0,5, correspondent à des vecteurs *X* non impliqués dans des relations de quasi-colinéarité. Les estimateurs MCO des coefficients de régression associés à ces vecteurs ne sont pas affectés par la multicolinéarité.

---

fait que les valeurs des variables explicatives doivent être ou non centrées (autour de leur moyenne) avant que l'estimation des indices de conditionnement et des proportions de décomposition des variances ne soit opérée (voir Erkel-Rousse, 1995, p. 34 et suiv.). Comme dans le cas des VIFs, Baltagi (2003, p.263) conseille d'opérer une analyse à partir des valeurs non centrées des variables explicatives, pour prendre en compte un impact possible de la constante, prise en compte que permet la commande « codiag2 ».

[5] Les lignes du tableau de décomposition des variances sont présentées selon une valeur croissante des indices de conditionnement.

[6] Selon Erkel-Rousse (1995), la valeur seuil de 30 est une valeur représentative d'une situation « aigüe » de multicolinéarité. En fait, un problème important peut être révélé dès lors qu'un ou plusieurs indices de conditionnement présentent une valeur supérieure ou égale à 20 selon cette auteure. Soulignons en outre qu'une situation de multicolinéarité « légère » est détectée dès lors que la valeur d'un indice de conditionnement est supérieure à 10, voire à 5 (BKW, 1980) …

- les indices de conditionnement présentant en revanche une valeur supérieure à 20 et présentant, pour chaque vecteur *X* et sur les lignes correspondant à ces indices et sur les lignes directement situées au-dessus dès lors que la valeur des indices de conditionnement correspondants dépasse 20, une somme des proportions de décomposition des variances supérieure à 0,5 ou 0,6 environ correspondent à des vecteurs *X* impliqués dans des relations de quasi-colinéarité. Les estimateurs MCO des coefficients de régression associés à ces vecteurs sont affectés par la multicolinéarité et ces coefficients peuvent avoir été estimés de façon très imprécise par leur estimateur MCO.

- enfin, plus la somme des proportions de décomposition des variances est proche de 1 et plus le diagnostic est préoccupant pour la précision des coefficients de régression. En outre, cette somme étant fixée, la précision des coefficients est d'autant plus faible que les proportions de décomposition des variances les plus élevées sont situées dans les lignes les plus basses du tableau de décomposition des variances. »

La règle énoncée ci-dessus permet de prévoir « toutes les configurations possibles de multicolinéarité » et donc de quantifier le nombre de dépendances proches existant entre les vecteurs *X*.

Un exemple est fourni au sein de l'article d'Erkel-Rousse (1995, p.29), dont nous reportons ici la teneur pour illustrer d'une façon pédagogique les propos qui viennent d'être tenus.

Imaginons un modèle RLC à 2 variables explicatives, X_1 et X_2, et imaginons que nous obtenions le tableau de décomposition des variances suivant à l'issue de l'estimation du modèle RLC :

Indices de conditionnement	Tableau de décomposition des variances		
	Var (*constante*)	Var (X_1)	Var (X_2)
1	0+	0+	0+
42	10^{-3}	10^{-3}	0,980
857	0,999	0,999	$2*10^{-1}$

Deux indices de conditionnement présentent des valeurs supérieures à 20. On est en présence d'une situation de multicolinéarité dite « d'ordre 2 ». L'examen de la ligne correspondant à l'indice de conditionnement le plus élevé - 857 - permet d'identifier une dépendance proche « dominante » entre la variable X_1 (dont la proportion de décomposition de la variance est 0,999, donc supérieure au seuil de 0,5) et la constante (qui présente une proportion de décomposition de la variance de 0,999, également supérieure au seuil de 0,5). Une dépendance proche « concurrente » est également identifiée dans un second temps puisque 1/ il existe un autre indice de conditionnement présentant une valeur supérieure à 20 - 42 - et 2/ la somme des proportions de décomposition des variances correspondant aux deux indices de conditionnement supérieurs à 20 ressort alors supérieure à 0,5 pour la variable X_2 (cette somme approchant de $1 = 0{,}980+2*10^{-1}$ pour X_2). Au final, l'analyse révèle donc l'existence de deux dépendances proches : une « dominante » entre X_1 et la constante et une « concurrente » associant X_2 avec la constante et X_1 (dont la multicolinéarité avait déjà été révélée au travers de la dépendance proche dominante).

Une fois qu'un phénomène de multicolinéarité a ainsi été révélé et que le nombre de dépendances proches a été quantifié, la seconde phase consiste à évaluer la portée de ce phénomène sur les résultats obtenus. BKW (1980) présentent un indicateur complémentaire pour ce faire, dénommé « signal-to-noise ratio » (et dont l'équivalent français pourrait être « ratio signal-bruit »). Ce ratio est calculé pour chacun des régresseurs et il permet d'identifier les variables explicatives qui auront « souffert » d'un problème de multicolinéarité. La méthode consiste à :

1/ calculer, une fois que le modèle RLC originel a été estimé, pour chacune des variables explicatives et pour la constante, la valeur de sa statistique *t* élevée au carré (t^2).

2/ comparer les statistiques t^2 obtenues à une valeur seuil identifiée au moyen de plusieurs tableaux élaborés à cet effet par Belsley (1982). Le choix de la valeur seuil dépend à la fois a/ du nombre de degrés de liberté du dénominateur, $n-p$ (n étant le nombre d'observations et p le nombre de régresseurs, constante incluse), et du nombre de degrés de liberté du numérateur, p_2 (p_2 étant le nombre de variables explicatives subissant le test), ayant servi à l'établissement des valeurs seuils b/ de la taille γ choisie pour le test (« test size ») et c/ du niveau d'adéquation (« level of adequacy ») α recherché pour le test. Selon Belsley (1984), la situation la plus classique consiste à considérer une taille de test γ de 0.999 et un niveau d'adéquation α de 0.05, ces deux valeurs correspondant aux modalités de choix d'une valeur seuil telles qu'elles sont présentées au sein du tableau figurant en Annexe 1 de l'article écrit en 1984 par ce même auteur (et auquel nous nous référerons par la suite). On obtient la valeur seuil en établissant en ligne la valeur $n-p$ et en colonne la valeur p_2. Par mesure de prudence, Douglass et al. (2003) conseillent d'opérer à partir d'une valeur p_2 égale à 1, ce qui correspond à la première colonne du tableau de l'article de 1984 (et donc de choisir la valeur seuil uniquement en fonction de la valeur de $n-p$).

Les régresseurs dont la valeur t^2 dépasse la valeur seuil identifiée n'ont pas été « touchés » par le problème de multicolinéarité et les statistiques *t* obtenues dans le cadre de l'estimation du modèle originel de régression sont interprétables comme dans le cas d'un modèle RLC non soumis à un problème de multicolinéarité ; en revanche, les variables explicatives dont la valeur t^2 est inférieure à la valeur seuil ont été « touchées » par ce phénomène.

Un problème de multicolinéarité existe donc à partir du moment où 1/ un phénomène de multicolinéarité a été mis en évidence (au moyen du tableau de décomposition des variances) 2/ ce phénomène introduit un ratio « signal-bruit » problématique pour un ou plusieurs régresseurs.

Les différentes étapes présentées permettent donc 1/ d'identifier un phénomène de multicolinéarité et 2/ de savoir quelle est la portée de ce phénomène sur les résultats obtenus, variable explicative par variable explicative, constante incluse.

Il peut donc apparaître étonnant que cette méthode n'ait pas été employée plus souvent.

Cela pourrait être lié au fait que cette méthode, telle qu'elle a été exposée par BKW (1980), apparaît complexe d'utilisation. En fait, elle peut être largement simplifiée à partir du moment où l'expérimentateur procède à la réalisation de tests univariés avant d'estimer un modèle RLC. En effet, l'existence d'un phénomène de multicolinéarité conduit à constater une forte dégradation de la significativité de la plupart des variables explicatives concernées. Les erreurs standard de ces variables ressortent alors avec une valeur importante, ce qui engendre à l'inverse une faible

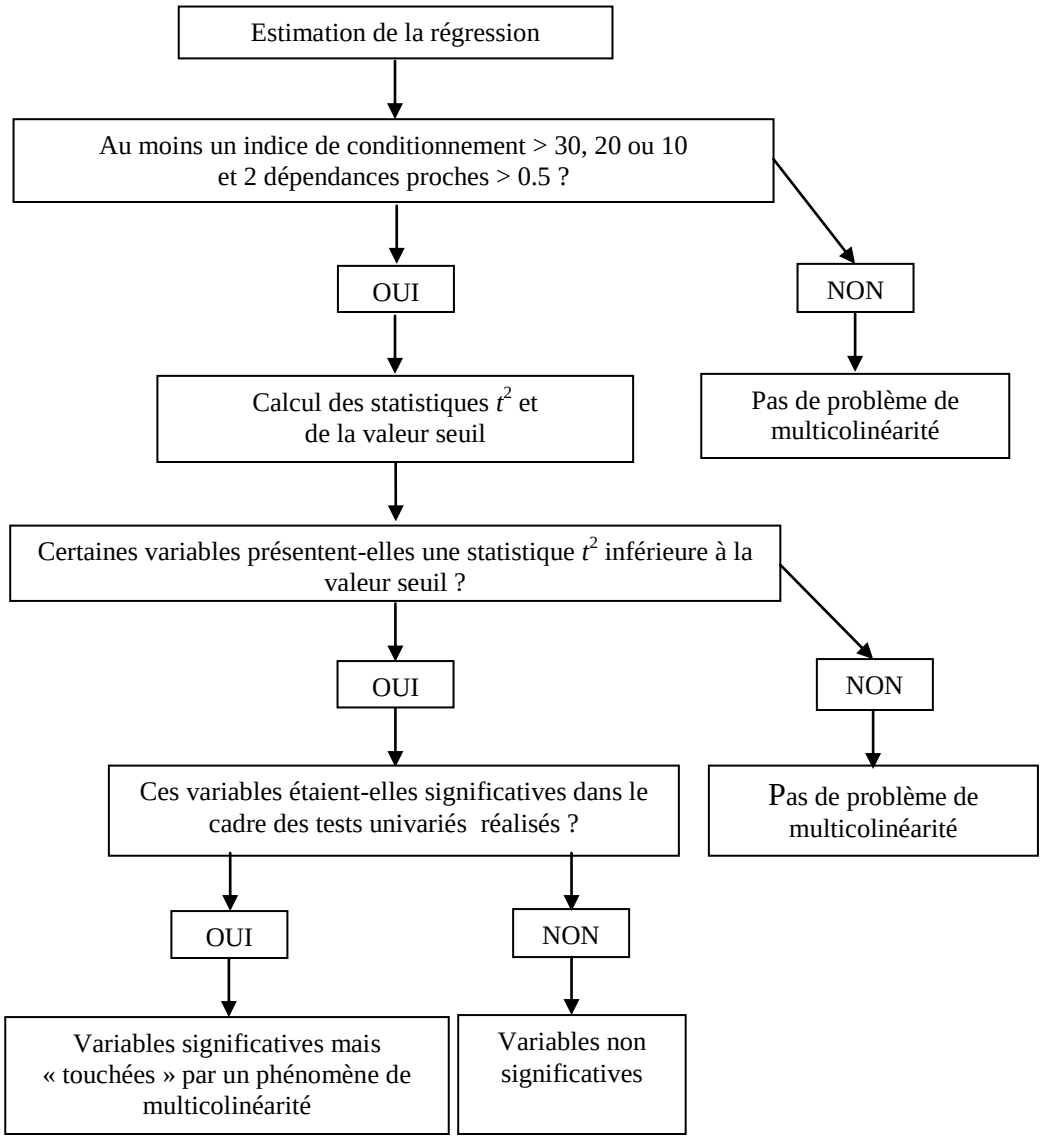
valeur des statistiques *t* recensées sur ces variables. On peut alors identifier les variables touchées par un problème de multicolinéarité en 1/ estimant suite aux tests univariés réalisés un modèle RLC à partir de l'ensemble des variables explicatives retenues pour l'analyse 2/ établissant le tableau de décomposition des variances 3/ calculant les statistiques t^2 et en les comparant à la valeur seuil établie. Les variables « touchées » par la multicolinéarité seront celles qui :

- étaient ressorties significatives (non significatives) dans le cadre des tests univariés et qui seront ressorties comme étant moins significatives / non significatives (faiblement significatives) dans le cadre de l'analyse multivariée.

- présenteront alors et systématiquement une statistique t^2 inférieure à la valeur seuil.

Ce cheminement peut être illustré par un algorithme de décision présenté au sein de la Figure 1 ci-dessous.

Figure 1 : Algorithme de décision simplifié relatif à la multicolinéarité



3. Etude empirique illustrative

Afin de montrer la praticité de l'algorithme proposé, une étude empirique est menée, qui porte sur les déterminants d'une publication volontaire d'informations sur les activités de R&D. Les développements qui suivent permettent 1/ de présenter les modalités de cette recherche 2/ d'énoncer les hypothèses relatives aux déterminants étudiés 3/ d'identifier le modèle RLC estimé 4/ d'apporter des précisions sur l'échantillon observé et 5/ de présenter les résultats à partir d'une lecture des indicateurs de BKW (1980), opérée selon l'algorithme de décision simplifié proposé.

3.1. Modalités de l'étude

L'étude s'intéresse aux déterminants d'une publication volontaire d'informations sur les activités de R&D, sur l'année 2002 (cet exemple est issu de travaux antérieurement réalisés dont l'intérêt était de présenter un problème important de multicolinéarité). Ce thème avait donné lieu à trois recherches antérieures, dont les résultats figurent au sein du Tableau 1.

Tableau 1 : Résultats des études antérieures

Auteurs	Pays	Année de traitement des données	Niveau de publication	Variables explicatives significatives et dans le sens positif attendu (1)	Variables explicatives non significatives
Entwistle (1999)	Canada	1994	Score de dénombrement	Intensité de R&D*** Cotation sur une place financière américaine*** Appartenance à un secteur d'activité innovant***	Inscription à l'actif des dépenses de R&D Taille Endettement
Percy (2000)	Australie	1993	Indice de publication	Intensité de R&D*** Financement étatique*** (2) Part non taxée des bénéfices** (2)	% Filiales non détenues à 100 % (2) Taille Endettement Coûts indirects Rentabilité Demande de capitaux
Ding et Stolowy (2003)	France	2000	Score de dénombrement	Cotation sur une ou plusieurs places financières anglo-saxonnes*** Appartenance à un secteur d'activité innovant*** Taille***	Intensité de R&D Inscription à l'actif des dépenses de R&D

(1) Résultats significatifs au seuil *** de 1% ; ** de 5 % ; * de 10 %
(2) Variables spécifiques au cas australien.

Au sein de notre étude, les niveaux de publication volontaire d'informations concernant les activités de R&D sont mesurés sous la forme d'un score de dénombrement (variable SCORE)[7], comme cela avait été fait par Entwistle (1999) et Ding et Stolowy (2003). Au sein de chaque rapport annuel des sociétés observées, les informations relatives aux activités de R&D ont été dénombrées (phrases et items quantitatifs), sachant que les mêmes règles de dépouillement des rapports annuels que celles Ding et Stolowy (2003) ont été adoptées. Ainsi :

« - Seules les informations relatives aux comptes consolidés ont été prises en compte ; les comptes sociaux sont hors du champ d'étude ;

- toute donnée chiffrée a été considérée comme une information. Néanmoins, quand le rapport annuel d'une entreprise présentait le montant brut et le montant net de la R&D immobilisée ainsi que le montant des amortissements, deux informations ont été comptabilisées et non trois (la troisième n'étant que la différence entre les deux premières).

- pour les données qualitatives, la phrase a été retenue comme l'unité d'information ; toute phrase ou toute proposition indépendante contenant une idée ou un ensemble d'idées cohérentes de même nature était donc comptabilisée comme une information. Néanmoins, lorsqu'une phrase contenait plusieurs idées sur la R&D, ce sont autant d'informations qui ont été comptabilisées. »

Et il a été décidé de procéder à l'étude de douze déterminants. Les hypothèses liées à ces douze facteurs explicatifs sont énoncées ci-après.

3.2. Hypothèses testées

Un examen approfondi des résultats des études portant sur les déterminants d'une publication volontaire d'informations a conduit à étudier douze variables explicatives. Sur ces douze variables :

- deux d'entre elles sont reliées aux activités de R&D (l'intensité de R&D et l'inscription à l'actif de tout ou partie des dépenses de R&D) ;

- l'une d'entre elles, la taille, peut être considérée comme une variable « multi-théories » ;

- une autre est représentative de l'importance des coûts d'agence (niveau d'endettement) ;

- une autre est représentative de l'importance des coûts politiques (appartenance à un secteur d'activité innovant) ;

- et trois d'entre elles peuvent être considérées comme apparentées à la théorie du signal (cotation sur une ou plusieurs places financières anglo-saxonnes, demande de capitaux autour de la période observée, volatilité boursière).

Il a été décidé d'étudier l'impact possible de trois autres variables, souvent testées au sein des recherches antérieures (renommée des auditeurs, degré d'internationalisation et niveau de

[7] A l'image de l'étude de Ding et Stolowy (2003), toute information publiée sur les activités de R& D est considérée ici comme ayant un caractère volontaire par nature, les textes français en vigueur à l'époque n'obligeant qu'à la publication d'informations « significatives » sur les activités de R&D, sans autre précision.

rentabilité), ainsi que celui d'une variable représentative de l'importance des coûts indirects (brevets).

L'intensité de R&D

Dans la mesure où cette étude s'intéresse aux déterminants d'une publication volontaire d'informations sur les activités de R&D, il est naturel d'émettre l'hypothèse que les rapports annuels d'entreprises fortement « intensives » en R&D (celles pour lesquelles les dépenses de R&D seront importantes par rapport au chiffre d'affaires) incorporeront davantage d'informations de nature volontaire sur ces activités que ceux d'entreprises faiblement « intensives » en R&D.

H1 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et l'intensité de R&D des entreprises observées.

Une politique d'inscription à l'actif de tout ou partie des dépenses de R&D

Pour ce qui concerne l'année d'observation retenue, les frais de recherche appliquée et de développement expérimental pouvaient, exceptionnellement, « être inscrits à l'actif sous réserve qu'un certain nombre de conditions soient simultanément remplies » (C.Com. art.D19, al.2 et PCG, art. 361-2) :

- les projets en cause devaient être nettement individualisés et leurs coûts distinctement établis pour pouvoir être répartis dans le temps ;

- chaque projet devait avoir une chance sérieuse de réussite technique et de rentabilité commerciale ;

- les dépenses de recherche et développement activées devaient être amorties dans un délai qui ne pouvait dépasser cinq ans.

Compte tenu de ces éléments, l'on pouvait s'attendre à ce que les rapports annuels d'entreprises au sein desquelles une politique d'inscription à l'actif de tout ou partie des dépenses de R&D était / avait pu être pratiquée incorporent davantage d'informations sur les activités de R&D que ceux des entreprises au sein desquelles un tel choix n'avait pas été / pas pu être opéré.

H2 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le fait que les dépenses de R&D des entreprises observées soient, pour tout ou partie, activées au bilan des ces entreprises.

La taille

La taille est un déterminant qui a été étudié d'une façon quasi systématique au sein des études portant sur les déterminants de la publication volontaire d'informations, pour plusieurs raisons. Premièrement, Jensen et Meckling (1976) ont montré analytiquement que les coûts d'agence augmentent en fonction de la taille des entreprises. Deuxièmement, plus « visibles » politiquement que les PME, les grandes entreprises sont généralement plus exposées à d'éventuels coûts politiques que les plus petites. Troisièmement, les grandes entreprises ont recours plus souvent que les PME aux marchés financiers. Les rapports annuels d'entreprises de taille importante devraient donc incorporer davantage d'informations de nature volontaire que ceux des plus petites entreprises.

H3 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et la taille des entreprises observées.

Le niveau d'endettement

Selon la théorie de l'agence (Jensen et Meckling, 1976), les coûts d'agence représentent une fonction croissante de la part de financement extérieur de l'entreprise. Les rapports annuels des entreprises les plus endettées devraient donc incorporer davantage d'informations de nature volontaire que ceux des entreprises les moins endettées.

H4 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le niveau d'endettement des entreprises observées.

L'appartenance à un secteur d'activité innovant

Selon la théorie politico-contractuelle, « les politiciens vont chercher à imposer des taxes supplémentaires aux entreprises dont les résultats seront élevés pour pouvoir redistribuer à leurs électeurs une partie de la richesse nationale sous forme de services publics gratuits, de subventions ou de tarifs protégés afin de se faire réélire » (Dumontier et Raffournier, 1999). Dans ce contexte, la publication volontaire d'informations apparaît comme un moyen d'éviter les coûts (politiques) qui seraient liés à ces actions. Les entreprises innovantes sont parfois susceptibles de dégager des résultats plus élevés que les entreprises opérant dans des secteurs traditionnels, ce qui les rend dans ce cas plus exposées à une intervention des pouvoirs publics. Les rapports annuels de ces entreprises devraient donc inclure davantage d'informations portant sur la R&D que ceux des entreprises opérant dans des secteurs traditionnels.

H5 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le fait que les entreprises observées appartiennent à un secteur d'activité innovant.

Une cotation sur une ou plusieurs places financières anglo-saxonnes

Les obligations de publication d'informations des sociétés cotées varient d'un pays à l'autre. Les rapports annuels des sociétés françaises cotées sur des places financières étrangères, et notamment sur des places financières anglo-saxonnes où les obligations de publication sont généralement plus importantes qu'en France, devraient donc incorporer davantage d'informations sur les activités de R&D que ceux des entreprises cotées sur une place financière française uniquement.

H6 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le fait que les titres des entreprises observées soient cotés sur un ou plusieurs marchés financiers anglo-saxons.

Une demande de capitaux sur les marchés financiers autour de la période d'analyse

Healy et Palepu (2001) postulent que les perceptions que les investisseurs ont d'une entreprise sont de toute première importance lorsque les dirigeants de cette entreprise prévoient une augmentation de capital ou la contraction d'un emprunt obligataire par l'intermédiaire des marchés financiers. Ainsi, on s'attend à une hausse du niveau de publication volontaire d'informations dans le cadre d'une demande de capitaux sur les marchés financiers avant, pendant ou après la période observée.

H7 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le fait que les entreprises observées aient procédé à une demande de capitaux sur les marchés financiers autour de la période d'analyse considérée.

La volatilité boursière

Les entreprises dont le cours boursier connaîtrait une volatilité importante pourraient être exposées, compte tenu de cette fluctuation, à un coût du capital important. On s'attend donc à ce que les rapports annuels d'entreprises dont le cours de Bourse connait d'importantes variations (ce qui peut être le cas des entreprises intensives en R&D, compte tenu d'un retour sur investissement pouvant se faire sur plusieurs années) incorporent de nombreuses informations de nature volontaire de façon à rassurer les investisseurs sur les perspectives de retour sur investissement liées aux activités de R&D.

H8 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et la volatilité du cours des actions des sociétés observées.

La « renommée » des auditeurs

Les entreprises « intensives » en R&D pourraient être exposées à des coûts d'agence et/ou à un coût du capital plus important que les sociétés opérant dans de secteurs d'activité « traditionnels ». On peut donc faire l'hypothèse que les dirigeants de ces entreprises recourront à des cabinets d'audit réputés afin de rassurer les créanciers, les actionnaires et les investisseurs. En contrepartie et afin de maintenir leur réputation, ces cabinets devraient inciter leurs clients à publier de nombreuses informations de nature volontaire sur les activités de R&D.

H9 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et la renommée des auditeurs des entreprises observées.

Le degré d'internationalisation

Plus les activités d'une entreprise sont diversifiées géographiquement, plus le niveau des informations publiées sur un mode volontaire au sein du rapport annuel de cette entreprise devrait être important. Ce phénomène s'expliquerait par « une nécessité des entreprises exportatrices de répondre ou de se conformer aux besoins d'information de leurs partenaires étrangers » (Bureau et Raffournier, 1989).

H10 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le degré d'internationalisation des entreprises observées.

Le niveau de rentabilité

Selon la théorie politico-contractuelle, les politiciens vont chercher à imposer des taxes supplémentaires aux entreprises dont les résultats seront élevés, engendrant ainsi des coûts « politiques » pour ces entreprises. Et, selon la théorie de l'agence, les dirigeants d'entreprises profitables pourraient être amenés à communiquer sur un mode volontaire à propos de ces résultats afin d'asseoir leur position et leur réputation suite à la réalisation de bonnes performances. Ces arguments impliquent que les niveaux de publication volontaire d'informations devraient être positivement associés au niveau de rentabilité des entreprises observées.

H11 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le niveau de rentabilité des entreprises observées.

La variable représentative de l'importance des coûts indirects (brevets)

Verrecchia (1983) a montré le rôle limitatif de la concurrence sur le processus de publication volontaire d'informations. Les coûts indirects liés à la publication volontaire d'informations sur

les activités de R&D pouvant être maîtrisés par une politique active de protection des innovations (brevets), l'hypothèse contraposée suivante est énoncée :

H12 : Il existe une association positive entre les niveaux de publication volontaire d'informations sur les activités de R&D et le fait qu'il existe une politique active de protection des innovations (brevets) au sein de ces entreprises.

Une présentation des douze déterminants étudiés est fournie au sein du Tableau 2. Ce tableau recense les définitions et dénominations opérationnelles qui ont été retenues pour chacun d'entre eux et le sens attendu de leur relation avec la variable dépendante.

3.3 Modèle de régession

En plus de la réalisation de tests univariés, le modèle de régression qui sera estimé par la suite se présente donc ainsi :

$$SCORE = \alpha + \beta_1 INTENS + \beta_2 COMPTA + \beta_3 TAILLE + \beta_4 DETTES + \beta_5 SECTEUR + \beta_6 COTAANG + \beta_7 DK + \beta_8 BETA + \beta_9 AUD + \beta_{10} CAEXP + \beta_{11} ROA + \beta_{12} BREVETS + \varepsilon$$

3.4. Informations relatives à l'échantillon

L'échantillon étudié est composé de l'ensemble des sociétés cotées sur le SBF 250 dont les rapports annuels incorporaient de l'information sur les activités de R&D en 2002. Après élimination des sociétés financières et immobilières en raison de leurs spécificités et élimination également de 7 sociétés présentant des valeurs extrêmes. 120 sociétés ont finalement été retenues en vue de la réalisation des tests statistiques.

Tableau 2 : Présentation résumée des variables explicatives

Hypothèse Testée	*Définition opérationnelle*	*Dénomination Opérationnelle*	*Sens attendu de la relation avec la variable dépendante*	*Source des données*
Intensité de R&D	Dépenses de R&D / Chiffre d'affaires	**INTENS**	+	Rapport Annuel
Inscription à l'actif de tout ou partie des dépenses de R&D	Variable dichotomique codée 1 si inscription pour tout ou partie des dépenses de R&D à l'actif du bilan et 0 dans le cas contraire	**COMPTA**	+	Rapport Annuel
Taille	Logarithme Total Bilan	**TAILLE**	+	Rapport Annuel
Niveau d'endettement	Dettes moyen long terme / (Total Bilan - Total des dépenses de R&D capitalisées)	**DETTES**	+	Rapport Annuel
Appartenance à un secteur d'activité innovant	Variable dichotomique codée 1 si l'entreprise observée appartient à un secteur d'activité innovant (Aéronautique/Défense ; Automobile ; Logiciel ; Matériel Informatique ; Pharmacie / Biotechnologies) et 0 dans le cas contraire (Consommation ; Industrie ; Ingéniérie ; Services)	**SECTEUR**	+	Rapport Annuel
Cotation sur une ou plusieurs places financières anglo-saxonnes	Variable dichotomique codée 1 si l'entreprise observée est cotée sur une ou plusieurs places financières anglo-saxonnes et 0 dans le cas contraire	**COTAANG**	+	Rapport Annuel
Demande de capitaux autour de la période observée	Variable dichotomique codée 1 si les dirigeants on procédé à une demande de capitaux sur les marchés en n-1, n, ou n+1 et 0 dans le cas contraire	**DK**	+	Rapport Annuel
Volatilité boursière	Beta de l'entreprise	**BETA**	+	Base Global Vantage
Renommée des auditeurs	Variable dichotomique codée 1 si les 2 auditeurs de l'entreprise observée font partie des « Big 4 » et 0 dans le cas contraire	**AUD**	+	Rapport Annuel
Degré d'internationalisation	% Chiffre d'affaires à l'exportation	**CAEXP**	+	Rapport Annuel
Niveau de rentabilité	(Résultat d'exploitation avant dotation aux amortissements et provisions – Montant R&D capitalisée sur l'année observée)/ (Total Bilan - Total des dépenses de R&D capitalisées)	**ROA**	+	Rapport Annuel
Coûts indirects	Variable dichotomique codée 1 s'il existe une politique affirmée de protection des innovations (brevets) et 0 dans le cas contraire	**BREVETS**	+	Rapport Annuel

3.5. Résultats de l'étude empirique

3.5.1. Statistiques descriptives

Le tableau 3 fournit les statistiques descriptives habituelles.

Tableau 3 : Statistiques descriptives

	n	***Moyenne***	***Médiane***	***Ecart-type***	***Minimum***	***Maximum***
SCORE	120	71	62	60	3	342
INTENS	103	0.046	0.025	0.061	0.00001	0.284
TAILLE (en M€)	118	9153	1160	18562	14	106584
DETTES	120	0.19	0.18	0.15	0.00	0.71
BETA	109	1.16	0.98	0.79	-0.13	3.17
CAEXP	114	0.61	0.66	0.25	0	1
ROA	118	0.00	0.02	0.13	-0.71	0.32
COMPTA	120	0.27		0.44	0	1
SECTEUR	120	0.39		0.49	0	1
COTAANG	120	0.31		0.46	0	1
DK	120	0.40		0.49	0	1
AUD	120	0.21		0.41	0	1
BREVET	120	0.59		0.49	0	1

Le niveau moyen de publication volontaire (SCORE) se situe à 71 informations relatives à la R&D. Ce chiffre est comparable à celui qui avait été obtenu au sein des études d'Entwistle (1999) et de Ding et Stolowy (2003). Cette situation moyenne doit être relativisée par d'importantes disparités inter-entreprises, le niveau minimum se situant à 3 items et le niveau maximum à 342 (avec un écart-type de 60), disparités que les auteurs des études précédentes avaient également relevé. Les statistiques qui concernent les variables indépendantes n'appellent pas de commentaires particuliers mais révèlent également (et cela rejoint aussi ce qui avait été constaté par les auteurs des études antérieures) l'existence de disparités importantes.

3.5.2. Résultats des tests univariés

Les douze hypothèses ont d'abord été testées sous une forme univariée. Des analyses préliminaires ayant révélé une absence de distribution gaussienne de l'ensemble des variables quatitatives (tests de Kolgomorov-Smirnov significatifs au seuil de 5 %), il a été décidé de transformer les valeurs de la variable dépendante SCORE et de la variable explicative TAILLE en prenant en compte le logarithme de ces variables, les valeurs ainsi transformées de la variable SCORE présentant alors une forme de distribution gaussienne[8] (la statistique *p* ressortant à 6% cela étant lors de la réalisation d'un nouveau test de Kolmogorov-Smirnov). Compte tenu de ce

[8] Cette transformation est habituelle lorsque la variable dépendante est une variable de « comptage » ne présentant pas à l'origine une forme de distribution gaussienne (voir par exemple l'étude de Piot, 2008, sur les délais d'audit).

dernier élément et du fait que les autres variables quantitatives n'aient pas pu être transformées de façon à approcher une forme de distribution gaussienne, il a été décidé de procéder à la réalisation de tests paramétriques et non paramétriques pour tester les douze hypothèses. Ainsi :

- les associations existant entre la variable SCORE et chacune des variables explicatives de nature quantitative ont été estimées à la fois au moyen de la technique des corrélations de Pearson (*r*) et de celle des corrélations de Spearman (*Rho*).

- et les associations existant entre la variable SCORE et chacune des variables explicatives dichotomiques ont été estimées à la fois au moyen de tests de Student (« t-tests ») - avec une correction appliquée en cas de variances inégales identifiées au moyen d'un test de Levene - et au moyen de tests de Mann-Withney (« z-tests »).

Les tableaux 4a et 4b reportent les résultats.

Leur lecture révèle l'existence d'une association positive, comme attendu et au seuil de 1 %, entre la variable SCORE et sept des douze déterminants examinés : l'intensité de R&D (INTENS), la taille (TAILLE), la part de chiffre d'affaires réalisée à l'exportation (CAEXP), l'appartenance à un secteur d'activité innovant (SECTEUR), une cotation du titre sur une ou plusieurs places financières anglo-saxonnes (COTAANG), une demande de capitaux autour de la période observée (DK) et l'existence d'une politique active de protection des innovations (BREVETS). La variable Niveau de rentabilité (ROA) ressort comme ayant un impact négativement significatif, contrairement à ce qui était attendu, à un seuil de 5 %, lorsque la méthode des corrélations de Spearman est appliquée et de 10 % seulement lorsque la méthode des corrélations de Pearson est appliquée. Les quatre autres variables - Niveau d'endettement (DETTES), Volatilité du titre (BETA), Inscription pour tout ou partie des dépenses de R&D à l'Actif (COMPTA) et Auditeur (AUD) sont non significatives.

Ces résultats viennent à la fois confirmer et compléter les résultats des trois études antérieures. Les entreprises dont les rapports annuels incorporent le plus d'informations de nature volontaire sur les activités de R&D sont ceux des entreprises :

- les plus grandes en termes de taille (comme cela avait été établi par Ding et Stolowy (2003))

- dont le titre est côté sur une ou plusieurs places financières anglo-saxonnes (comme cela avait également été établi par Ding et Stolowy (2003))

- et qui appartiennent à un secteur d'activité innovant (comme cela avait été établi par Entwistle (1999) et Ding et Stolowy (2003))

En outre, l'analyse montre une absence de significativité des variables Inscription de tout ou partie des dépenses de R&D à l'Actif (à l'image des études d'Entwistle (1999) et Ding et Stolowy (2003)) et de la variable Niveau d'endettement (à l'image de l'étude de Percy (2000)).

Pour ce qui est des éléments complémentaires, l'analyse révèle une significativité positive au seuil de 1 % entre l'Intensité de R&D et les niveaux de publication sur les activités de R&D, ce qui vient confirmer les résultats des études d'Entwistle (1999) et Percy (2000) mais est en contradiction avec les résultats non significatifs établis sur cette variable dans le cas français (Ding et Stolowy, 2003). Après enquête, ce phénomène était lié à une différence de composition

d'échantillon entre les deux études. Une analyse ad hoc, à composition d'échantillons égale, a finalement permis d'établir un résultat significativement positif sur cette variable pour les deux études, ce qui a permis de statuer sur un impact significativement positif de cette variable dans le cas français, rejoignant ainsi les résultats obtenus au sein des deux autres études antérieures[9].

Notre analyse révèle également l'existence d'une association significativement positive entre les niveaux de publication volontaire sur les activités de R&D et :

- la Part de chiffre d'affaires réalisée à l'exportation. Ce résultat vient renforcer le résultat significativement positif qui avait été établi sur cette même variable au sein de l'étude de Depoers (1999), qui portait sur les niveaux généraux de publication volontaire au sein des rapports annuels dans le cas français.

- le Niveau de rentabilité. Il semble donc que cette variable ait un impact négatif sur les niveaux de publication sur la R&D dans le cas français alors qu'elle n'avait aucun impact dans le cas australien (Percy, 2000).

Enfin, ces résultats révèlent, dans le contexte français, une absence de relation significative entre les niveaux de publication volontaire sur les activités de R&D et :

- la Volatilité du titre des entreprises observées.

- la Renommée de leurs auditeurs. Ce dernier résultat vient confirmer une absence de significativité de cette variable lorsque la publication d'informations qui est à l'étude porte sur des activités stratégiques (à l'image du résultat non significatif obtenu sur cette variable par Craswell et Taylor (1992) pour leur étude portant sur la publication volontaire d'informations sur les réserves énergétiques).

[9] De fait, nos travaux ont également mis en évidence l'intérêt de réaliser une étude portant sur les déterminants d'une publication volontaire d'informations sur plusieurs années et non une seule. Nous remercions à cet effet les Professeurs Ding et Sotolwy d'avoir bien voulu nous fournir leurs données concernant l'année 2000.

Tableau 4a : Tests univariés (variables quantitatives)

	r	***Rho***
INTENS	**0.414*** (0.000)**	**0.526*** (0.000)**
TAILLE	**0.353*** (0.000)**	**0.384*** (0.000)**
DETTES	-0.081 (0.381)	-0.125 (0.241)
BETA	0.059 (0.541)	0.028 (0.794)
CAEXP	**0.356*** (0.000)**	**0.299*** (0.004)**
ROA	**-0.176* (0.057)**	**-0.231** (0.028)**

Tableau 4b : Tests univariés (variables dichotomiques)

		Valeur de Score (en nombre d'informations)		
		1	***0***	***Test Student (t)/ Mann-Whitney (z)***
COMPTA	Moy.	74.0	69.6	0.54
	Méd.	54.5	63.0	0.11
SECTEUR	Moy.	100.9	51.4	**5.36*** **
	Méd.	83.0	40.0	**4.79*** **
COTAANG	Moy.	111.8	52.6	**5.81*** **
	Méd.	94.0	40.0	**5.05*** **
DK	Moy.	88.6	58.9	**2.68*** **
	Méd.	74.5	45.5	**2.66*** **
AUD	Moy.	80.2	68.3	1.03
	Méd.	67.0	57.0	0.89
BREVETS	Moy.	94.4	36.7	**7.20*** **
	Méd.	78.0	25.0	**6.20*** **

**** ,**,* relation significative au seuil de 1, 5 et 10 % respectivement*

3.5.3 Résultats de l'analyse multivariée

Il a ensuite été estimé un modèle RLC. A l'issue de l'estimation, il a été procédé à la vérification du respect des onze conditions nécessaires à la validité d'un tel modèle, au moyen des tests décrits au sein de l'Annexe 1. Toutes les hypothèses, à l'exception de l'Hypothèse 11, étaient respectées. Le seul problème était donc celui de l'existence d'un phénomène de multicolinéarité. Les résultats sont présentés au sein du tableau 5, qui fournit également les valeurs des VIFs (centrés et non centrés), les tableaux 6 et 7 permettant quant à eux de fournir les valeurs :

- des coefficients de corrélation (et de leurs niveaux de signification) obtenus suite à la réalisation d'une matrice des corrélations de Pearson ;

- des indicateurs de BKW - tableau de décomposition des variances -.

Tableau 5 : Résultats de la régression

	Coef.	***T***	***P>t***	***VIFs (centrés)***	***VIFs (non centrés)***
INTENS	**3.93**	**2.87**	**0.005*** **	2.53	3.95
COMPTA	**0.25**	**1.83**	**0.071***	1.39	1.91
TAILLE	**0.20**	**5.20**	**0.000*** **	2.38	**258.90**
DETTES	-0.33	-0.67	0.504	1.57	4.79
SECTEUR	**0.58**	**3.79**	**0.000*** **	2.08	3.39
COTAANG	0.02	0.12	0.905	1.84	2.79
DK	0.09	0.74	0.460	1.42	2.36
BETA	-0.03	-0.42	0.675	1.60	5.00
AUD	-0.19	-1.47	0.147	1.18	1.55
CAEXP	0.03	0.13	0.899	1.44	**9.98**
ROA	-0.98	-1.55	0.125	1.20	1.25
BREVETS	**0.82**	**6.44**	**0.000*** **	1.47	3.31
Constante	-1.18	-1.44	0.153		**246.58**
Moyenne VIFs				1.68	**41.98**
Sig.			*0.0000*		
R^2			*0.7536*		
N			*88*		

***,**,* relation significative au seuil de 1 %, 5 % et 10 % respectivement

Tableau 6 : Matrice des corrélations

	INTENS	*COMPTA*	*TAILLE*	*DETTES*	*SECTEUR*	*COTAANG*	*DK*	*BETA*	*AUD*	*CAEXP*	*ROA*	*BREVETS*
INTENS	1.000											
COMPTA	-0.059	1.000										
	(0.585)											
TAILLE	-0.270**	-0.112	1.000									
	(0.011)	(0.299)										
DETTES	-0.254**	-0.043	0.206*	1.000								
	(0.017)	(0.693)	(0.055)									
SECTEUR	**0.653*** **	0.195	**-0.314*** **	-0.141	1.000							
	(0.000)	(0.068)*	**(0.003)**	(0.190)								
COTAANG	0.258	-0.225	**0.476*** **	0.038	0.069	1.000						
	(0.015)**	(0.035)**	**(0.000)**	(0.729)	(0.521)							
DK	-0.067	-0.081	**0.419*** **	**0.303*** **	-0.073	**0.297*** **	1.000					
	(0.537)	(0.456)	**(0.000)**	**(0.004)**	(0.501)	**(0.005)**						
BETA	**0.346*** **	0.297***	**-0.306*** **	0.084	**0.353*** **	0.049	0.003	1.000				
	(0.001)	(0.005)	**(0.004)**	(0.437)	**(0.001)**	(0.649)	(0.977)					
AUD	-0.023	0.016	**0.292*** **	-0.047	-0.116	0.104	-0.019	0.007	1.000			
	(0.828)	(0.880)	**(0.006)**	(0.665)	(0.283)	(0.337)	(0.859)	(0.950)				
CAEXP	0.221**	-0.211**	**0.333*** **	0.039	-0.005	**0.405*** **	**0.292*** **	0.045	0.169	1.000		
	(0.039)	(0.049)	**(0.002)**	(0.715)	(0.967)	**(0.000)**	**(0.006)**	(0.679)	(0.116)			
ROA	-0.096	-0.177*	-0.040	-0.245**	-0.108	-0.043	-0.058	-0.184*	-0.006	0.025	1.000	
	(0.375)	(0.099)	(0.713)	(0.022)	(0.319)	(0.688)	(0.590)	(0.086)	(0.955)	(0.815)		
BREVETS	0.219**	-0.173	0.169	**-0.322*** **	0.144	**0.304*** **	0.117	-0.166	0.070	0.249**	-0.054	1.000
	(0.040)	(0.108)	(0.115)	**(0.002)**	(0.180)	**(0.004)**	(0.276)	(0.123)	(0.516)	(0.020)	(0.619)	

***,**,* relation significative au seuil de 1 %, 5 % et 10 % respectivement
(…) Sig. Les relations significatives au seuil de 1 % sont indiquées en gras.

Tableau 7 : Tableau de décomposition des variances

Vecteur	*Indice de Conditionnement*	*Constante*	*INTENS*	*COMPTA*	*TAILLE*	*DETTES*	*SECTEUR*	*COTAANG*	*DK*	*BETA*	*AUD*	*CAEXP*	*ROA*	*BREVETS*
1	1.00	0.000	0.002	0.003	0.000	0.003	0.003	0.003	0.004	0.003	0.003	0.002	0.000	0.003
2	2.50	0.000	0.019	0.079	0.000	0.001	0.043	0.014	0.020	0.007	0.023	0.002	0.226	0.003
3	2.69	0.000	0.068	0.071	0.000	0.025	0.042	0.012	0.019	0.000	0.030	0.000	0.115	0.010
4	2.81	0.000	0.002	0.119	0.000	0.001	0.003	0.071	0.034	0.004	0.005	0.000	0.380	0.009
5	3.11	0.000	0.004	0.000	0.000	0.017	0.001	0.001	0.069	0.001	0.644	0.000	0.027	0.005
6	4.02	0.000	0.022	0.319	0.000	0.092	0.003	0.040	0.025	0.030	0.029	0.000	0.001	0.204
7	4.20	0.001	0.007	0.091	0.000	0.005	0.004	0.235	0.122	0.008	0.039	0.005	0.051	0.228
8	4.51	0.000	0.010	0.018	0.000	0.008	0.063	0.302	0.545	0.012	0.102	0.002	0.000	0.006
9	5.76	0.000	0.122	0.000	0.000	0.147	0.494	0.052	0.036	0.251	0.025	0.012	0.006	0.000
10	6.69	0.000	0.406	0.159	0.000	0.010	0.131	0.008	0.000	0.537	0.006	0.044	0.004	0.110
11	8.25	0.002	0.268	0.081	0.002	0.555	0.195	0.000	0.002	0.007	0.010	0.238	0.152	0.359
12	10.24	0.013	0.019	0.048	0.011	0.136	0.012	0.017	0.041	0.008	0.001	0.680	0.026	0.062
13	**60.55**	**0.984**	0.051	0.012	**0.986**	0.002	0.007	0.243	0.083	0.133	0.090	0.015	0.012	0.000

Les résultats font apparaître une significativité positive de cinq variables explicatives : les variables INTENS, TAILLE, SECTEUR et BREVETS au seuil de 1 % et la variable COMPTA, au seuil de 10 % (qui était pourtant non significative dans le cadre des tests univariés). La significativité constatée dans le cadre des tests univariés pour les variables COTAANG, DK, CAEXP et ROA a quant à elle disparue.

Ces écarts de résultats sont liés au problème de multicolinéarité évoqué, qui est mis en lumière par les outils de diagnostic classiques, à l'exception cependant des VIFs centrés, qui auraient conduit à conclure qu'il n'existait pas de problème de multicolinéarité. On se rend donc ici compte de l'importance d'estimer les VIFs non centrés.

Une analyse opérée à partir de ces derniers révèle en effet un problème de multicolinéarité existant entre les variables TAILLE (VIF de 258,90), CAEXP (VIF proche de 10) et la constante de régression (VIF de 246,58), la même analyse opérée à partir de l'examen de la matrice des corrélations de Pearson révélant quant à elle un problème multiple de multicolinéarité (voir les relations présentant des coefficients de corrélation supérieures à 0,3 au sein du Tableau 6).

Les outils classiques de détection (à l'exception des VIFs centrés) révèlent donc bien un problème de multicolinéarité. Néanmoins, ils ne permettent pas de statuer quant à la portée de ce problème sur les résultats obtenus.

Un examen des indicateurs de BKW est plus informatif à cet égard. Il permet de constater l'existence d'une dépendance proche « dominante » entre la variable TAILLE et la constante. Les valeurs de référence des indicateurs de BKW étant dépassées, on est bien en présence d'un problème de multicolinéarité. En considérant une valeur seuil de 5 pour les indices de conditionnement et en appliquant la règle RI d'Erkel-Rousse (1995), quasiment toutes les autres variables explicatives sont concernées par des dépendances proches « concurrentes », à un plus faible degré de multicolinéarité cependant.

Afin de résoudre ce problème, l'algorithme de décision qui a été présenté a alors été appliqué. Les statistiques t^2 ont été calculées et la valeur seuil a été identifiée afin de savoir quelles étaient les variables qui avaient été « touchées » par la multicolinéarité. La lecture du tableau de Belsley (1984) a permis d'identifier cette valeur, qui est égale à 25.78 (pour $n = 88$, $p = 13$, $p_2 = 1$, $\gamma = 0.999$ et $\alpha = 0.05$). Il apparait donc que l'ensemble des régresseurs sont « touchés » par le problème de multicolinéarité, à l'exception des variables TAILLE et BREVETS (Tableau 8). Une comparaison des résultats obtenus dans le cadre des tests univariés avec ceux de l'analyse multivariée permet alors de statuer définitivement quant à la significativité des variables explicatives.

Les variables qui ressortent significatives à l'issue de l'analyse globale des tests réalisés sont les suivantes :

- TAILLE et BREVETS, significativement positives tant dans le cadre des tests univariés que dans le cadre de l'analyse multivariée (et qui n'ont donc pas été touchées par le phénomène de multicolinéarité identifié).

- INTENS et SECTEUR, significativement positives tant dans le cadre des tests univariés que dans le cadre de l'analyse multivariée mais qui ont été touchées par le phénomène de multicolinéarité identifié.

- COTAANG, DK, CAEXP, significativement positives dans le cadre des tests univariés réalisés mais non significatives dans le cadre de l'analyse multivariée réalisée. Ces variables ont donc été touchées plus sévèrement que les variables INTENS et SECTEUR par le phénomène de multicolinéarité identifié.

- la variable ROA, significativement négative dans le cadre des tests univariés réalisés mais non significative dans le cadre de l'analyse multivariée réalisée. Cette variable a donc été « touchée » par le phénomène de multicolonéarité identifié.

Les variables COMPTA, DETTES, BETA et AUD, non significatives dans le cadre des tests univariés, peuvent quant à elles être définitivement considérées comme non significatives.

Les conclusions qui avaient émané des tests univariés restent donc valables mais une analyse multivariée a pu être réalisée en maitrisant le problème de multicolinéarité rencontré, ce qui a permis de mettre en évidence une influence prépondérante des variables INTENS, TAILLE, SECTEUR et BREVETS sur les niveaux de publication sur les activités de R&D par rapport aux autres variables explicatives étudiées.

Tableau 8 : Statistiques t^2 et décision finale

Variable	t^2	$t^2 <$ Valeur seuil ?	**Variable potentiellement touchée par la multicolinéarité**	Variable significative dans le cadre de l'analyse multivariée ?	Variable significative dans le cadre des tests univariés ?	**Variable significative ?**
INTENS	8.24	O	**O**	O	O	**O**
COMPTA	3.35	O	**O**	O	N	**N**
TAILLE	27.04	N	**N**	O	O	**O**
DETTES	0.45	O	**O**	N	N	**N**
SECTEUR	14.36	O	**O**	O	O	**O**
COTAANG	0.01	O	**O**	N	O	**O**
DK	0.55	O	**O**	N	O	**O**
BETA	0.18	O	**O**	N	N	**N**
AUD	2.16	O	**O**	N	N	**N**
CAEXP	0.02	O	**O**	N	O	**O**
ROA	2.40	O	**O**	N	O	**O**
BREVETS	41.47	N	**N**	O	O	**O**

O : Oui N : Non

Conclusion

Cet article avait pour objet de s'intéresser au traitement d'un problème de multicolinéarité au sein des études portant sur les déterminants d'une publication volontaire d'informations.

Après avoir présenté les problèmes liés à l'existence d'un tel phénomène et rappelé quelles étaient les solutions habituellement utilisées pour remédier à ces problèmes, il a été montré en quoi les indicateurs proposés par Belsley, Kuh et Welsch (1980) apparaissaient comme des outils appropriés pour détecter un problème de multicolinéarité et identifier la portée de ce problème sur les résultats obtenus.

A partir de ces indicateurs, un algorithme de décision simplifié a été proposé, qui a été appliqué au sein d'une étude empirique illustrative portant sur les déterminants d'une publication volontaire d'informations sur les activités de recherche et développement.

Cette étude, qui a permis tout à la fois de confirmer et de compléter les résultats des études antérieures, a également permis de démontrer en quoi l'algorithme de décision proposé pouvait permettre de maitriser les problèmes liés à l'existence d'un phénomène de multicolinéarité.

Cet algorithme pourrait ainsi s'avérer d'une utilité certaine dans le cadre de la réalisation d'autres études de nature quantitative en Comptabilité.

Bibliographie

Ahmed, K. Courtis, J. K. (1999). Associations between corporate characteristics and disclosure levels in annual reports: a meta-analysis. *British Accounting Review* 31 (1): 35-61.

Akerlof, G. (1970). The market for lemons: qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics* 89: 488-500.

Baltagi, B.H. (2003). *A companion to theoretical econometrics*. Malden: Blackwell Publishing.

Belsley, D.A. (1982). Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics* 20: 211-253.

Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting* 3: 183-196.

Belsley, D. A., Kuh, E. Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.

Bourbonnais, R. (2005). *Econométrie*. Editions Dunod (6ème edition).

Bureau, D., Raffournier, B. (1989). Some determinants on accounting choices for consolidated statements of French firms: the case of pension costs engagements. *Papier présenté au 12ème Congrès de l'EAA, Stuttgart*.

Chatterjee, S., Hadi, A.S., Price, B. (2000). *Regression analysis by example*. John Wiley & Sons.

Chavent, M., Ding, Y., Fu, L., Stolowy, H., Wang, H. (2006). Disclosure and determinants studies: An extension using the divisive clustering method (DIV). *European Accounting Review* 15(2): 181-218.

Condoers, G., Saez, M. (2000). Collinearity, Heteroscedasticity and Outlier Diagnostics in Regression. Do They Always Offer What They Claim? *New Approaches in Applied Statistics* 16: 79-94.

Craswell, A., Taylor, S. (1992). Discretionary disclosure of reserves by oil and gas companies: an economic analysis. *Journal of Business Finance and Accounting* 19: 295-308.

Depoers, F. (1999). Contribution à l'analyse des déterminants de l'offre volontaire d'information des sociétés cotées. Thèse pour le doctorat de troisième cycle, Université Paris IX Dauphine.

Ding, Y., Stolowy, H. (2003). Les facteurs déterminants de la stratégie des groupes français en matière de communication sur les activités de recherche et développement. *Finance-Contrôle-Stratégie* 6 (1): 39-62.

Douglass, D.H., Clader, B.D., Christy, J.R., Michaels, P.J., Belsley, D.A. (2003). Test for harmful collinearity among predictor variables used in modeling global temperature. *Climate Research* 24: 15-18.

Dumontier, P., Raffournier, B. (1999). Vingt ans de recherche positive en comptabilité financière. *Comptabilité - Contrôle - Audit*, 5: 179-197.

Entwistle, G. M. (1999). Exploring the R&D disclosure environment. *Accounting Horizons* 13 (4): 323-341.

Erkel-Rousse, H. (1995). Détection de la multicolinéarité dans un modèle linéaire ordinaire : quelques éléments pour un usage averti des indicateurs de Belsley, Kuh et Welsch. *Revue de Statistique Appliquée* 43 (4): 19-42.

Garcia-Meca, E., Sanchez-Balesta, J.P. (2010). The association of board independence and ownership concentration with voluntary disclosure. *European Accounting Review* 19 (3): 603-627.

Gujarati, D.N. (2004). *Econométrie*. Editons de Boeck (4ème edition).

Healy, P. M., Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. *Journal of Accounting and Economics* 31: 405-440.

Jensen, M., Meckling, W. (1976). Theory of the firm: managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics* 3 (4): 305-360

Meek, G. K., Gray, S.J, Roberts, C. B. (1995a). Factors influencing voluntary annual report disclosures by U.S., U.K. and continental European multinational corporations. *Journal of International Business Studies* 26 (3): 555-572.

Percy, M. (2000). Financial reporting discretion and voluntary disclosure: corporate research and development expenditure in Australia. *Asia-Pacific Journal of Accounting and Economics* 7: 1-31.

Piot, C. (2008). Les déterminants du délai de signature du rapport d'audit en France, *Comptabilité-Contrôle-Audit* 14(2): 43-74.

Pourtier, F. (2004). La publication d'informations financières volontaires : synthèse et discussions. *Comptabilité- Contrôle-Audit* 10 (1): 79-102.

Verrecchia, R.E. (1983). Discretionary disclosure. *Journal of Accounting and Economics* 5: 195-211.

Watts, R. (1977). Corporate financial statements, a product of the market and political processes. *Australian Journal of Management* 2: 53-75.

Watts, R., Zimmerman, J. (1978). Towards a positive theory of the determination of accounting standards. *The Accounting Review* 53 (1): 112-134.

Watts, R., Zimmerman, J. (1986). *Positive Accounting Theory*. Prentice Hall.

Annexe 1 :
Procédures à utiliser sous STATA 11 pour tester du respect des hypothèses

Hypothèses	**Tests (« ligne de commande à entrer sous STATA une fois que le modèle originel a été estimé »)**	**Solution(s) si hypothèse non respectée**
1 : Nombre d'observations > nombre de régresseurs	Hypothèse supposée respectée	Un certain nombre d'économètres recommande même que le ratio (Nombre d'observations / Nombre de régresseurs) soit supérieur à 5 pour fournir des estimations utiles des paramètres β, *F et t*
2 : Linéarité	Linéarité (« **linktest** »)	1/ Transformation des variables quantitatives (sous la forme de logarithmes, de racines carrées, de puissances…), qui aboutira à « linéariser » la relation existant entre la variable à expliquer et les variables explicatives 2/ Choix d'autres estimateurs que les MCO (régressions non linéaires)
3 : Modèle RLC correctement spécifié	Bonne spécification (« **ovtest** »)	1/ Choix d'autres régresseurs en complément des régresseurs étudiés 2/ Choix d'autres estimateurs que les MCO)
4 : Valeurs fixes des régresseurs	Hypothèse supposée respectée	
5 : Terme d'erreur et *X* non corrélés	Calcul de corrélations entre le terme d'erreur et les *X*	Méthode des variables instrumentales
6 : Valeur moyenne de l'erreur est nulle		Selon Gujarati (2004), le respect de cette hypothèse n'est crucial que si la constante a une importance, ce qui n'est que rarement le cas. On ignorera cette hypothèse en général.
7 : Homoscédasticité	- Breusch et Pagan (« **hettest** ») - White (« **whitetst** ») *NB* : Des études récentes (Coenders et Saez, 2000, notamment) ayant révélé une hypersensibilité du test de Breusch et Pagan en cas de problème d'aplatissement et/ ou de non-normalité concernant la forme de distribution du terme d'erreur, le test de White sera préféré.	Autres méthodes d'estimation (qui vont dépendre de la forme d'hétéroscédasticité rencontrée) : 1/ Moindres Carrés Généralisés (MCG) 2/ Moindres Carrés Pondérés (MCP) 3/ Régression de type « robuste » …

Annexe 1 (Suite et fin) :

8 : Distribution normale du terme d'erreur	Kolgomorov-Smirnov (« **sktest** *résidus* ») *NB* : le respect de cette hypothèse n'est important que dans le cas de petits échantillons (< 100 observations). Elle n'a donc pas à être testée lorsque le nombre d'observations dépasse 100.	Le non-respect de cette hypothèse est le plus souvent lié à l'existence d'observations éloignées ou extrêmes (« outliers »). Il convient de repérer ces observations, au moyen : 1/ soit d'une boite à moustaches réalisée pour chacune des variables quantitatives 2/ soit de différentes commandes post-régression (calcul des résidus studentisés et/ou standardisés, distance de Cook, distance de Welsch, méthode du levier, DFBETA, COVRATIO, DFITs…)
9 : Absence d'autocorrélation des erreurs	Durbin-Watson NB : l'autocorrélation se rencontre essentiellement dans les modèles en séries temporelles ou longitudinaux. Le respect de cette hypothèse est de ce fait rarement testé avec des modèles en coupe instantanée, qui sont ceux qui sont concernés prioritairement par cet article.	Régression à la Prais-Winsten
10 : Stabilité suffisante prise par les régresseurs	Pas de test spécifique ; le non-respect de cette hypothèse se traduit par un faible nombre de variables significatives. Ce phénomène est en général lié soit à un trop faible nombre d'observations, soit à un choix inadéquat des variables étudiées soit à un phénomène de multicolinéarité	- Agrandir la taille d'échantillon - Choisir d'autres variables explicatives - Remédier au problème de multicolinéarité